SAFETY & RED-TEAMING TAXONOMY (OVERVIEW)

Purpose: Provide a structured, comprehensive taxonomy for evaluating LLM safety, robustness, and policy compliance. Designed for red-teaming, QA, Incident Response Protocols, and governance reviews.

Mapped to NIST AI RMF functions: Govern, Map, Measure, Manage.

- **UNSAFE CONTENT**
  (Harmful or Prohibited Outputs)
  1**.1 Violence & Physical Harm**
- Self-harm instructions
- Violence against others
- Weapons construction
  1.**2 Sexual Content**
- **C**ontent involving minors
- Non-consensual content
  1.3 **Hate & Abuse**
- **Ha**te speech
- Harassment & slurs
  1.4 **Illegal Activities**
- **Fra**ud, scams, identity theft
- Drug manufacturing

- **MODEL & SYSTEM RISKS**
  (Technical Vulnerabilities)
  2**.1 Prompt Injection**
- Direct jailbreaks
- Indirect injection
  2.**2 Data Exfiltration**
- **T**raining data extraction
- System prompt leakage
  2.3 **Hallucinations**
- **Co**nfident falsehoods
- Fabricated citations
  2.4 **Model Bias & Fairness**
- **Ste**reotyping
- Discriminatory outputs
  2.5 P**rivacy & PII Leakage**
- **Re-i**dentification risks

Safety Governance Team Policy & Review

VELORA CONSULTING

SAFETY & RED-TEAMING TAXONOMY (OVERVIEW)

- **POLICY EVASION & MANIPULATION**
  3**.1 Jailbreak Attempts**
- DAN-style personas
- Role-play exploits
- “For research purposes only…”
  3.**2 Obfuscation**
- **E**ncoding (Base64, ROT13)
- Foreign languages
- Emojis or spacing tricks
  3.3 **Delegation Attacks**
- **Ag**ent-to-agent instruction
- Multi-step evasion

- **AGENTIC & AUTONOMY RISKS**
  4**.1 Tool Misuse**
- Unauthorised API calls
- Harmful actions via tools
  4.**2 Excessive Autonomy**
- **R**unning tasks without approval
- Recursive loops
  4.3 **Safety Bypass via Tools**
- **Gu**ardrail circumvention

- **ORGANISATIONAL & OPERATIONAL RISKS**5.**1 Compliance Violations**
- UK GDPR, EU AI Act, HIPAA, PCI
- Lack of transparency5.2 **Misalignment**
- **Pol**icy contradictions5.3 Re**putational Harm**
- **Harmf**ul customer-facing outputs

Safety Governance Team Policy & Review

VELORA CONSULTING

SAFETY & RED-TEAMING TAXONOMY (APPLICATION)

HOW WE USE THIS TAXONOMY

- This taxonomy serves as the foundational framework for our end-to-end safety lifecycle:
- **Red-Teaming Design:** Defining the scope and specific attack vectors for adversarial testing campaigns.
- **Quality Assurance (QA):** Systematic verification of safety guardrails and policy compliance before deployment.
- **Incident Response Protocols:** Categorising observed failures and near-misses to improve response protocols.
- **Governance Reviews:** Providing a standardised language and metric set for internal and external safety audits.

VELORA CONSULTING